

PDEs, I/O, and Performance Visualization

Bill Gropp

gropp@mcs.anl.gov

www.mcs.anl.gov/~gropp



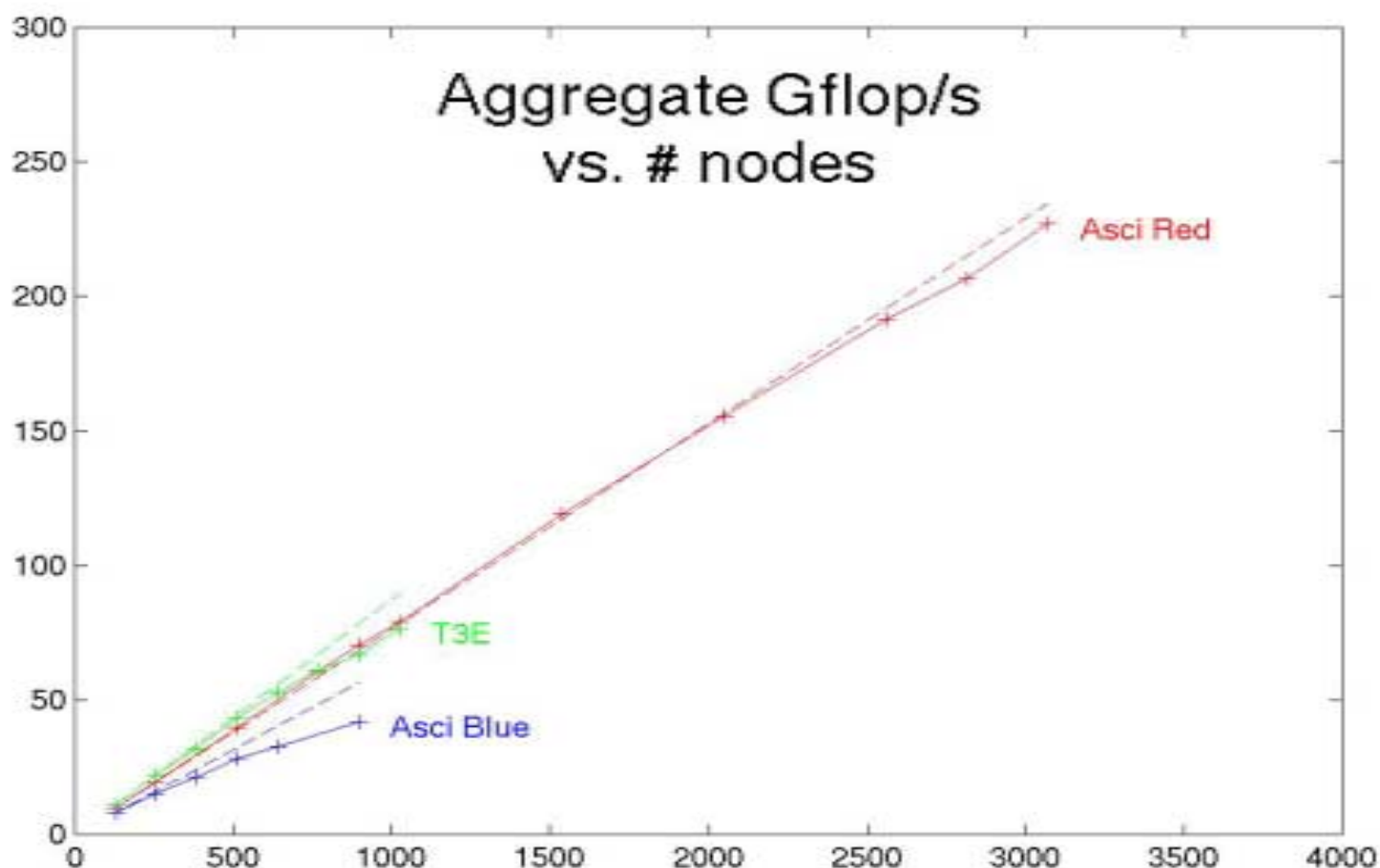
Parallel Solution of PDEs with PETSc and Tao

- www.mcs.anl.gov/petsc
 - ◆ Linear and nonlinear solvers, timestepping, distributed data management
 - ◆ Wide variety of algorithms, integrated performance tools
- www.mcs.anl.gov/tao
 - ◆ Unconstrained and bound-constrained optimization.
 - ◆ Nonlinearly constrained problems under development

Background of PETSc

- Developed by Gropp, Smith, McInnes & Balay (ANL) to support research, prototyping, and production parallel solutions of operator equations in message-passing environments
- Distributed data structures as fundamental objects—index sets, vectors/gridfunctions, and matrices/arrays
- Iterative linear and nonlinear solvers, combinable modularly and recursively, and extensibly
- Portable, and callable from C, C++, Fortran
- Uniform high-level API, with multi-layered entry
- Aggressively optimized: copies minimized, communication aggregated and overlapped, caches and registers reused, memory chunks preallocated, inspector-executor model for repetitive tasks (e.g., gather/scatter)
- Supports a wide variety of sparse matrix formats, including user-defined.
- Extensible with user-defined preconditioners, iterative methods, etc.

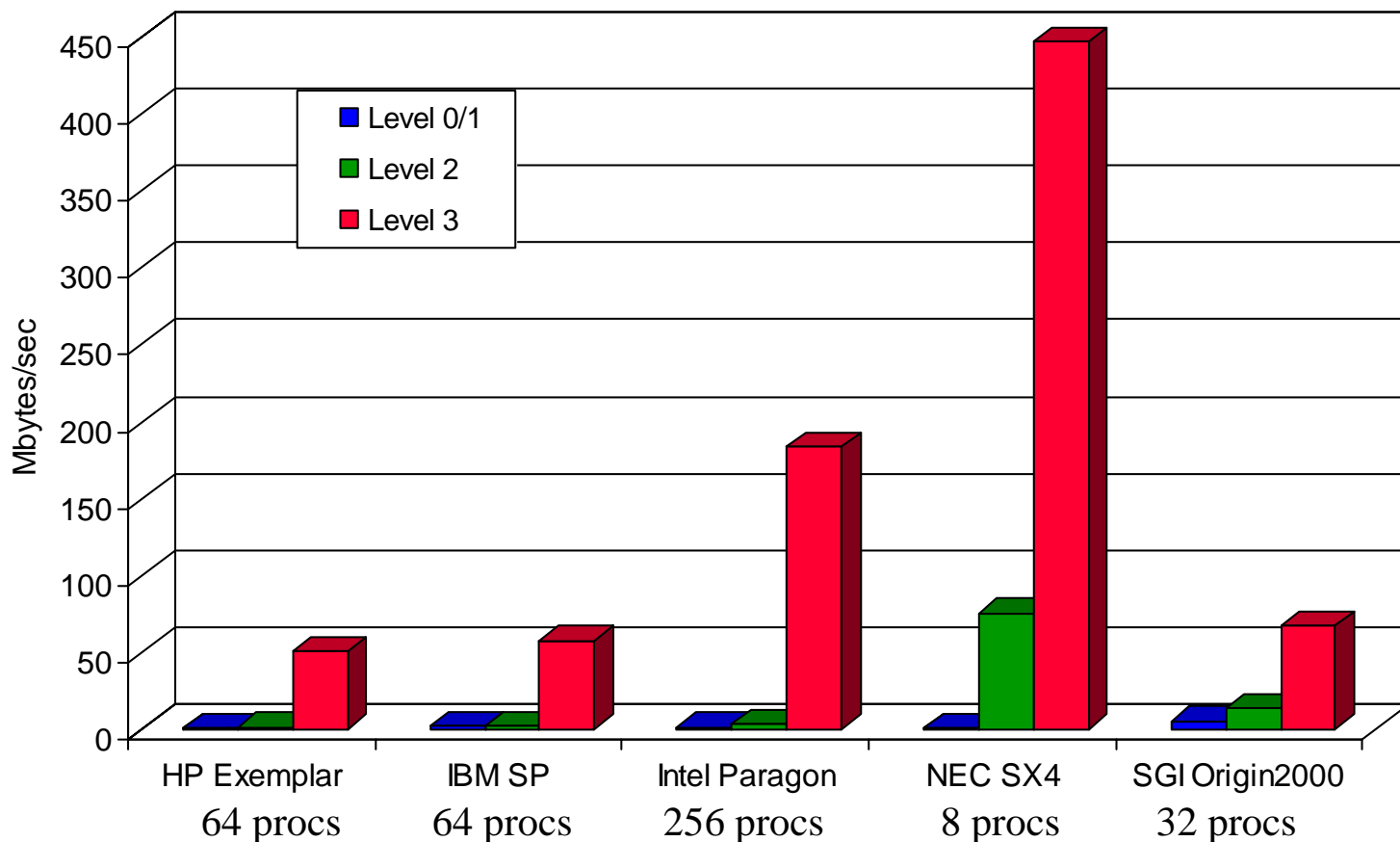
Fixed-size Parallel Scaling Results for Fun3d (Unstructured Fortran CFD code)



Parallel I/O

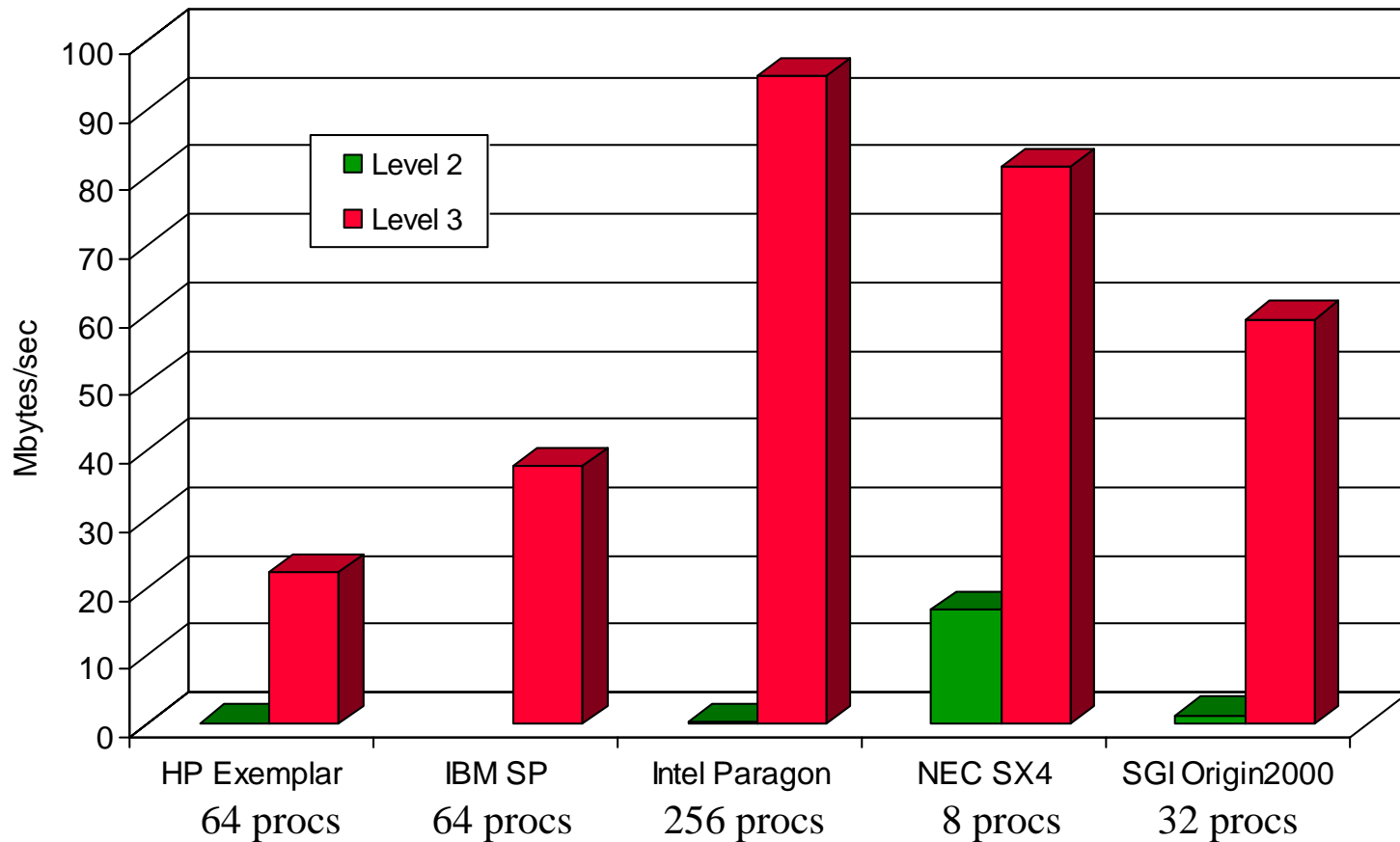
- MPI-IO
 - ◆ Standards-based parallel I/O
- PVFS: Parallel Virtual File System
 - ◆ www.parl.clemson.edu/pvfs
 - ◆ Already running on IA64 and mixed IA64/IA32 (at OSC)
- Wide-area file transfers
 - ◆ Use non-stream semantics to optimize transfers (non-TCP)

Distributed Array Access: Write Bandwidth

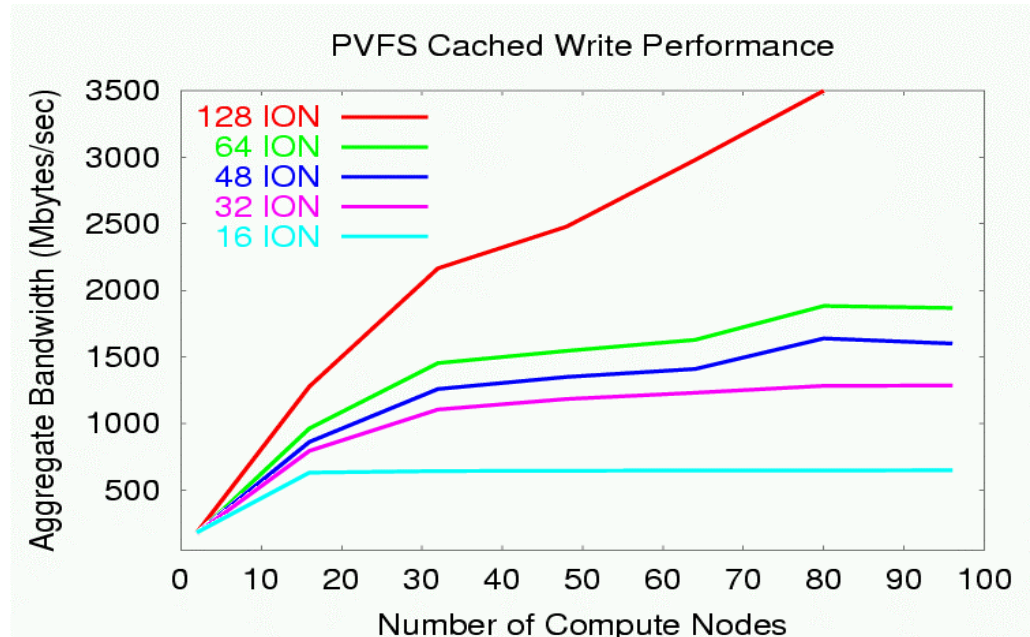


Array size: 512 x 512 x 512

Unstructured Code: Write Bandwidth

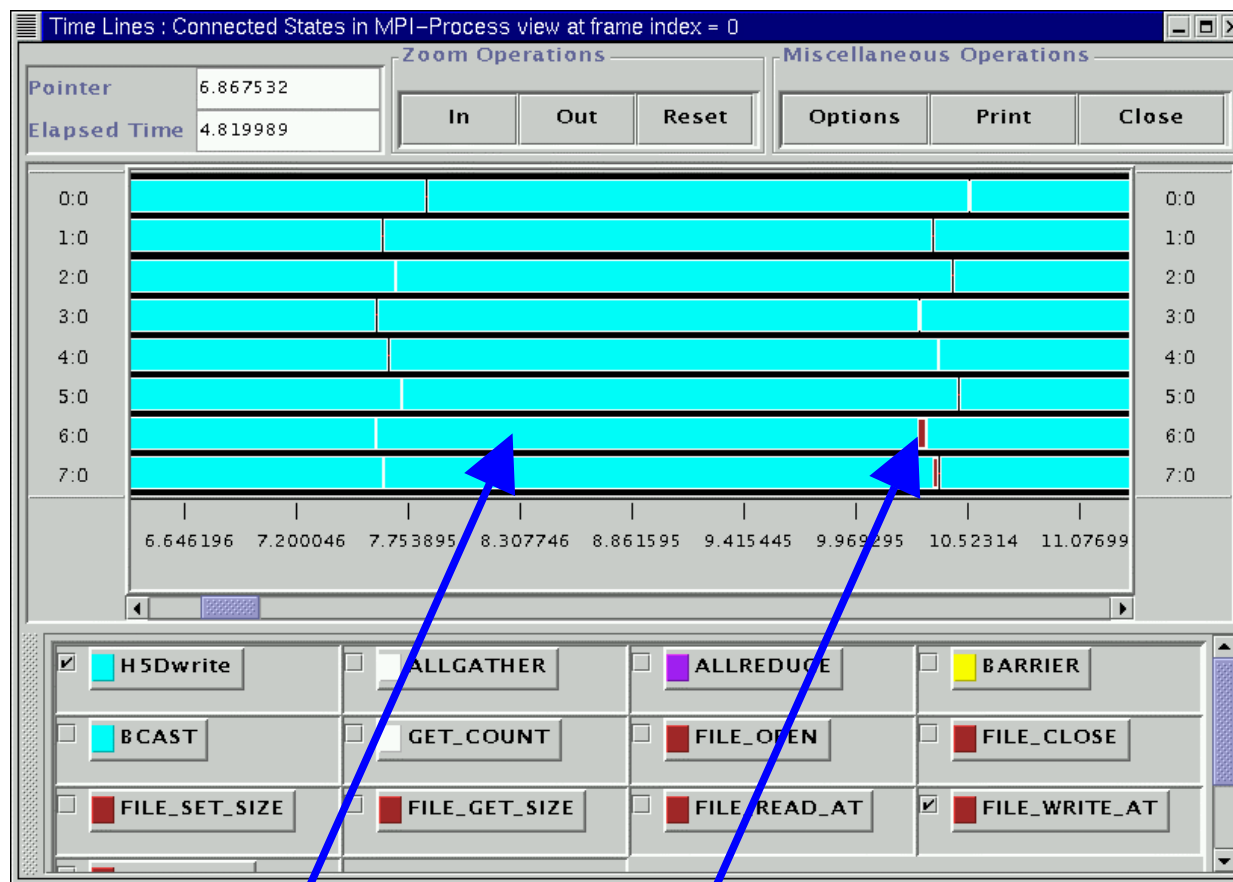


PVFS Peak Write Performance



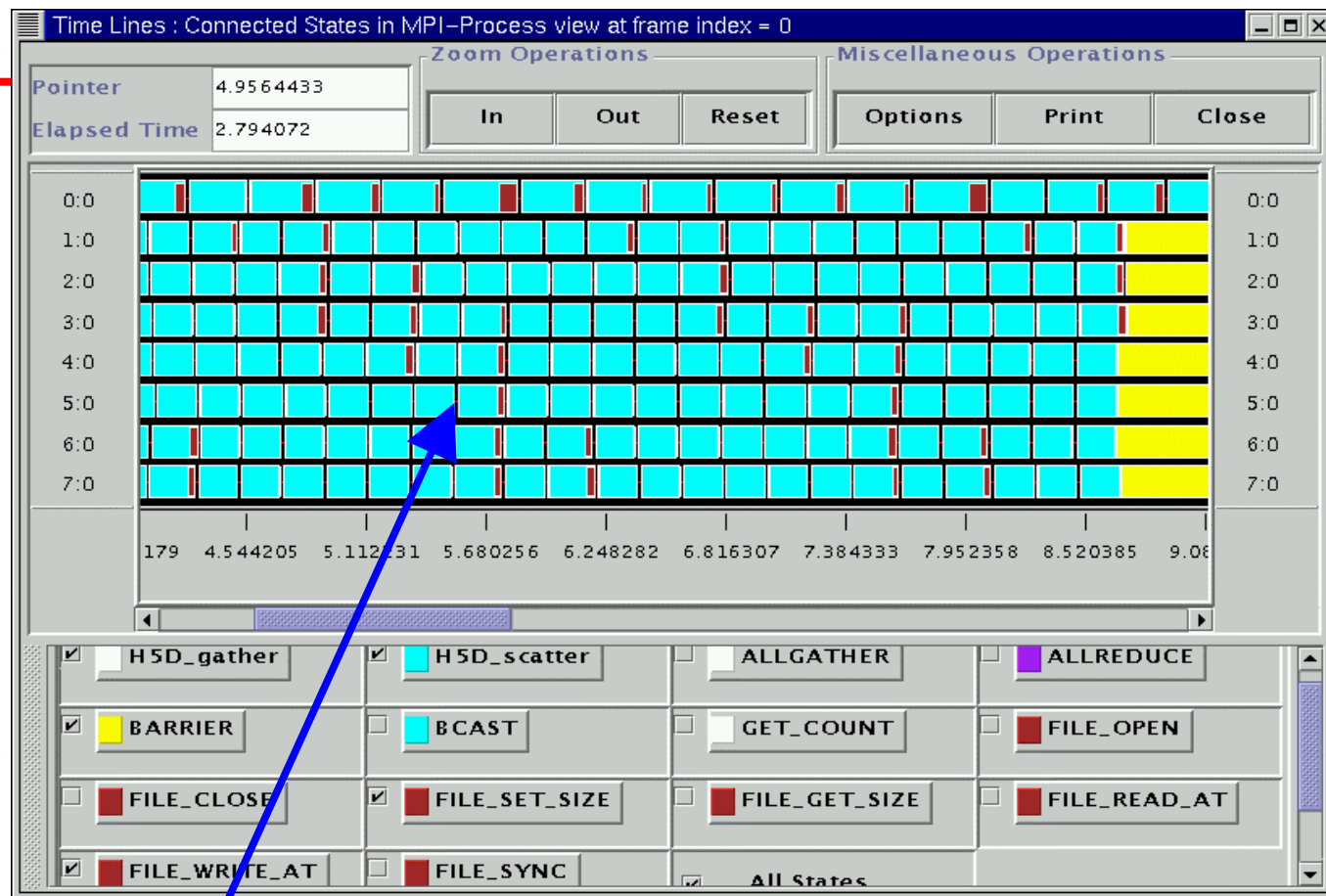
- Using compute nodes for storage in these tests
- Peak at around 25-30 Mbytes/sec per I/O server
- Clients cannot maintain this to disk

Jumpshot State View - H5Dwrite times



H5Dwrite MPI_File_write_at

Flash I/O benchmark with hand-coded packing



- Now with 64 processors, see 41.7 MBytes/sec

Parallel NetCDF

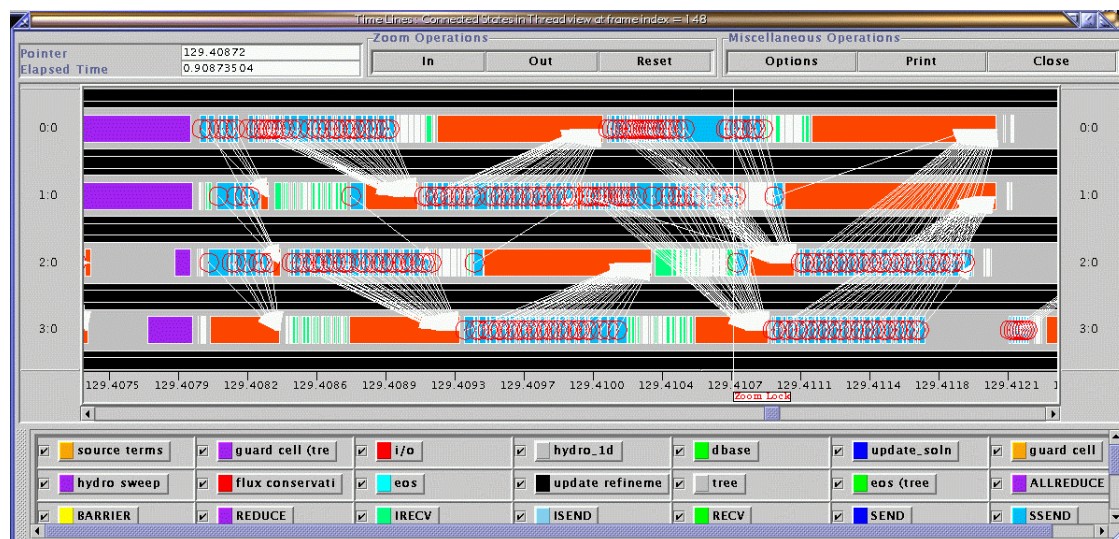
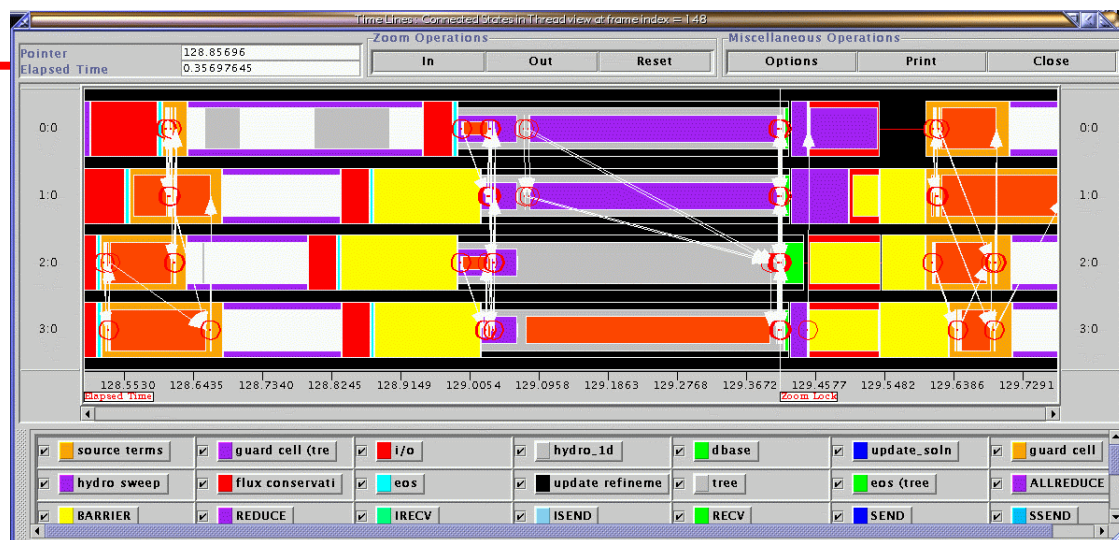
- NetCDF is a common file format and API for storage of scientific data
 - ◆ Sequential design limits performance
- Maintain the file format for portability
- Present new API for parallel access to datasets
- Provide flexibility for underlying implementation
- Build on top of MPI-IO
 - ◆ Supported on numerous platforms/file systems
 - ◆ ROMIO optimizations for free

Performance Understanding and Visualization

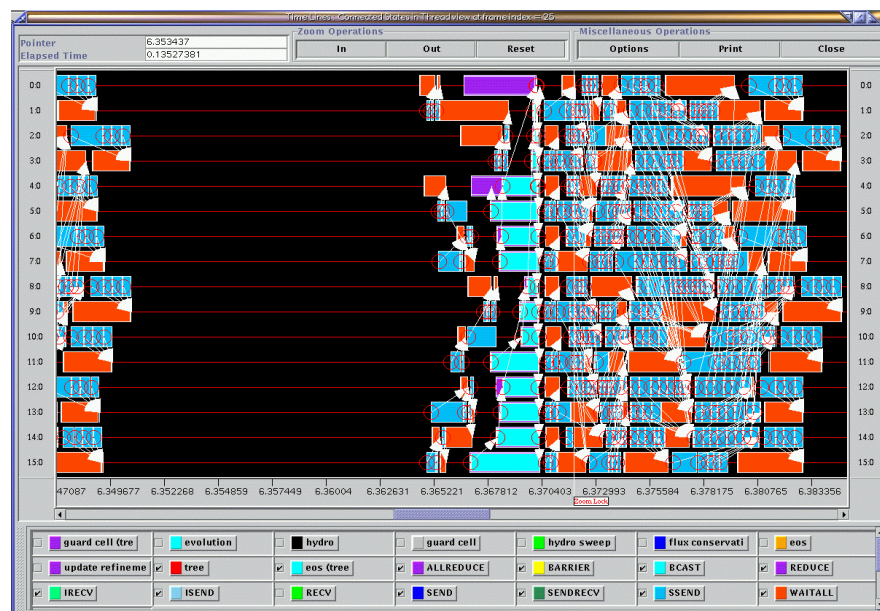
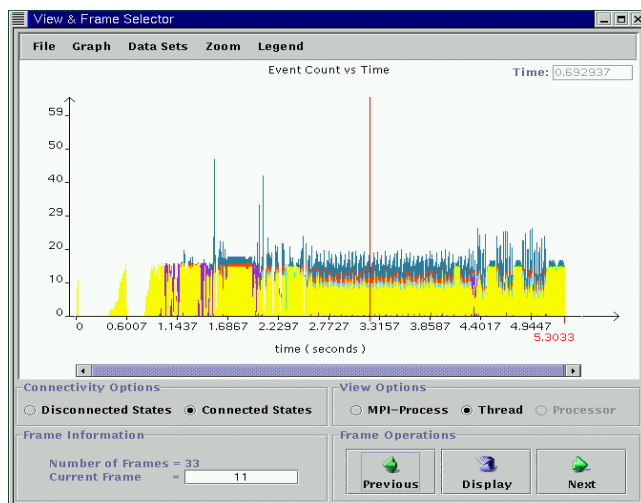
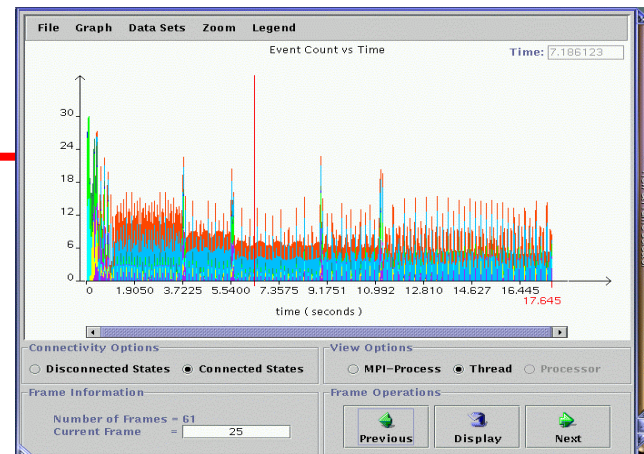
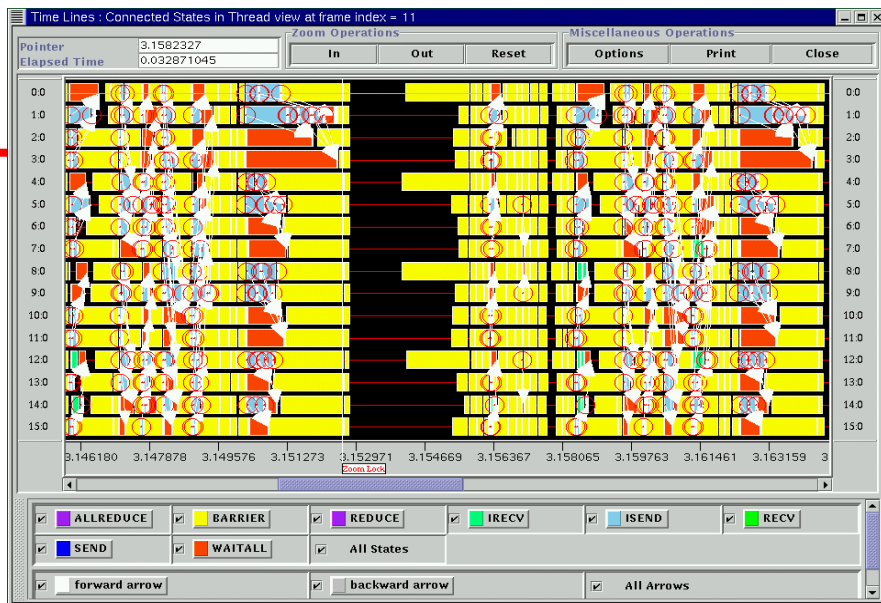
- Exploit the MPI profiling interface to gather data
 - ◆ Requires only that programs be *relinked*; no source changes or recompilation required
- Slog and Jumpshot
 - ◆ Detailed visualization of every MPI operation
 - ◆ Scalable handling of enormous log files (>1GB)
 - ◆ www.mcs.anl.gov/perfvis
- FPMPI
 - ◆ Summary data organized by message length and partner, along with access to performance counters
 - ◆ Multiple output formats including XML for interoperation with other tools

Using Jumpshot

- MPI functions and messages automatically logged
- User-defined states
- Nested states
- Zooming and scrolling
- Spotting opportunities for optimization



Removing Barriers From Paramesh



Sample output from FPMPI trace run

Program /u/ac/gunter/Cactus/exe/cactus_einstein
MPI_Finalize called at: Wed Nov 07 10:40:00 2001

Number of processors: 196

Timing Stats: [seconds] [min/max] [min rank/max rank]
wall-clock: 961.7 sec 836.540000 / 987.950000 70 / 156
user: 602.1 sec 534.486408 / 724.564124 7 / 129
sys: 359.9 sec 197.551578 / 451.912724 45 / 158

Memory Usage Stats (RSS) [min/max KB]: 203536/215808

----- MPI Routine Statistics -----

----- Barriers and Waits -----

MPI_Wait 14472 14472 14472 0 0 5.377592 73.969772 45.319035 93 98

----- Message Routines -----

	[max bin limit]	[min/max/avg calls]	[min rank/max rank]	[min/max/avg msg size]	[min rank/max rank]	[min/max/avg time]	[min rank/max rank]
MPI_Allreduce	: 8	45 45 45	0 0	360 360 360	0 0	0.203895 0.243965 0.225373	0 95 7.666142
MPI_Isend	: 64 K	3618 7236 5943	0 33	141825600 283651200 232999200	0 33	1.014190 5.456218 2.972432	3 105
MPI_Irecv	: 64 K	3618 7236 5943	0 33	141825600 283651200 232999200	0 33	0.068625 4.718235 1.721471	195 90

----- Message Totals -----

Immediate send calls: 3618 7236 5943 0 33

Immediate send size: 141825600 283651200 232999200 0 33

Immediate send time: 1.014190 5.456218 2.972432 3 105

Immediate comm rate: 47990746 139841307 82704196 110 3

Collective Send calls: 45 45 45 0 0

Collective Send sizes: 360 360 360 0 0

Collective Send times: 0.203895 0.243965 0.225373 0 95

Collective Comm rate: 1475.622713 1765.612923 1600.155819 95 0

Collective sync times: 7.666142 452.316393 159.509933 157 71

----- Hardware Performance Stats

Event 0 156723247838

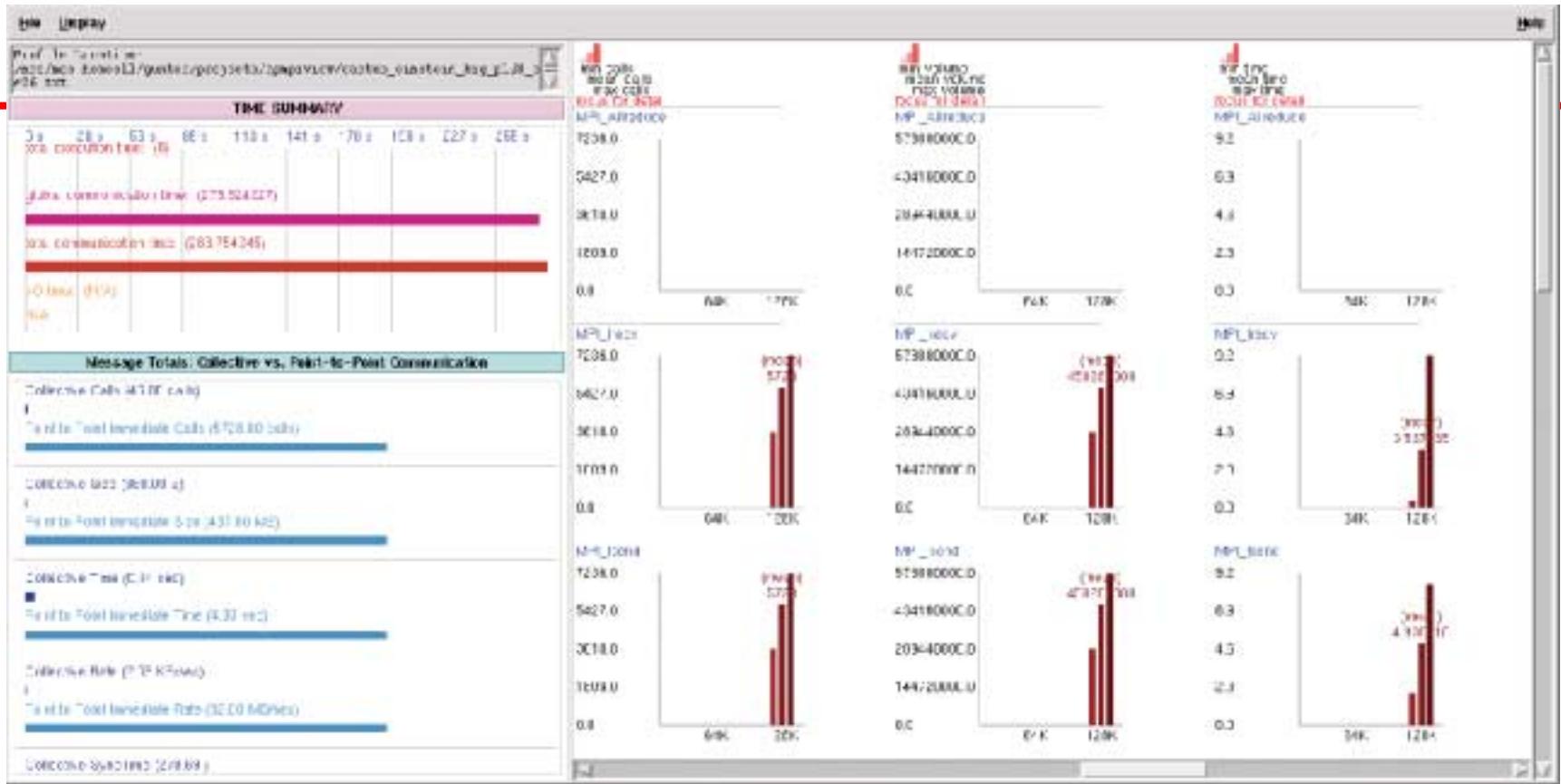
Event 26 493735494

----- Communication Partners -----

MPI_Isend 3 6 4.9 0 33

MPI_Irecv 3 6 4.9 0 33

FPMPI View



- ◆ <http://www.ncsa.uiuc.edu/TechFocus/Projects/NCSA/PerfSuite.html>